

D4.7: An Ethical framework for the development and use of AI and robotics technologies

[WP4: Artificial Intelligence and Robotics - Ethical, Legal and Social Analysis]

Lead contributor	Prof. dr. Philip Brey, University of Twente p.a.e.brey@utwente.nl
Other contributors	Philip Jansen, University of Twente Jonne Maas, University of Twente Björn Lundgren, University of Twente Anaïs Resseguier, Trilateral Research
Reviewers	Laura Crompton Bernd Carsten Stahl
Commentator	Rowena Rodrigues
Due date	31 March, 2020
Delivery date	8 April 2020 (after request for small extension)
Type	Dissemination Public
Dissemination level	PU = Public
Keywords	Artificial Intelligence; robotics; ethical framework; ethics by design; research ethics; policy; education; standards

© SIENNA, 2020

This work is licensed under a Creative Commons Attribution 4.0 International License

The SIENNA project - *Stakeholder-informed ethics for new technologies with high socio-economic and human rights impact* - has received funding under the European Union's H2020 research and innovation programme under grant agreement No 741716.



Abstract

This report proposes a comprehensive strategy for ethical AI and robotics. That is, it proposes, at least in outline, a comprehensive set of methods and procedures for developing, deploying and using AI and robotics systems in a way that adheres to ethical principles. The strategy that we propose addresses all actors in society, particularly developers, deployers, users, regulators and educators. It proposes various methods towards more ethical development and use of AI and robotics, such as methods for incorporating ethical considerations into design and development processes, guidelines for ethical deployment and use of AI and robotics systems, standards and certification, governmental policies and regulations, and education and training programs. We conclude this report by looking forward to the steps that still need to be taken to further develop and implement our strategy.

Document history

Version	Date	Description	Reason for change	Distribution
V0.9	01 03 2020	Final draft for external review	-	01 03 2020
VX.X	08 04 2020	Final report for submission to the EC	Reviews and comments	08 04 2020
V2.0	16 12 2020	Revised version	Removed Ethics by Design approach, will appear in D5.4	16 12 2020

Information in this report that may influence other SIENNA tasks

Linked task	Points of relevance
D5.4	The code of responsible conduct for AI and robotics will require consideration of the issues identified in this report.
D6.1	The report on adapting methods for ethical analysis of emerging technologies will require contemplation about the successes and challenges in the methodology used to write this report.
D6.3	The step-by-step guidance from ethical analysis to ethical codes and operational guidelines task will require reflection about the successes and challenges in writing this report.
D6.4	The process of obtaining buy-in for the codes from EU and international institutions will need to build on the proposals in this report.

Table of contents



Abstract 2

Table of contents 3

Executive summary 4

List of figures 6

List of tables 6

List of acronyms/abbreviations 6

Glossary of terms 6

1. Introduction 7

 1.1 Background 7

 1.2 Objectives 7

 1.3 Structure of the report 7

 1.4 Scope and limitations 8

2. A Strategy for Ethical AI and Robotics 10

3. Conclusion 22

References 24

Submitted to H2020, awaiting approval



Executive summary

This report proposes a comprehensive strategy for ethical AI and robotics. That is, it proposes, at least in outline, a comprehensive set of methods and procedures for developing, deploying and using AI and robotics systems in a way that adheres to ethical principles.

The report contains an introductory section, in which the objectives, scope and limitations of the report are set out, two main sections in which our strategy is presented, and finally a concluding section and two annexes. The two main sections of the report are sections 2 and 3. Section 2, “A Strategy for AI and Robotics,” proposes the overall strategy for promoting ethical AI and robotics. It is stated that a strategy for ethical AI and robotics should contain three components: (1) an identification of relevant actors; (2) an identification of methods that these actors can use to contribute to ethical AI & robotics, and (3) proposals of ways in which these methods can be made available to these actors, and ways to motivate them to use them. Following this proposal, the report continues to identify main classes of relevant actors who can bring about ethical AI and robotics: AI & robotics developers; AI & robotics development support organizations; organizations that deploy and use AI & robotics technology; governance and standards organizations; educational and media organizations; and civil society organizations and the general public.

Next, six types of methods for ethical AI & robotics are discussed and related to these classes of actors:

1. Methods for incorporating ethics into research and development of AI & robotics (aimed at AI & robotics developers and support organizations). These methods include research ethics guidelines and protocols for R&I in AI & robotics, ethical impact assessment methodologies for emerging AI & robotics, Ethics by Design methodologies for AI & robotics and codes of professional ethics for researchers and developers of AI & robotics technologies.
2. Methods for incorporating ethics into the deployment and use of AI & robotics (aimed at organisations that deploy and use AI & robotics technology). These methods include operational ethics guidelines and protocols for the deployment and use of AI and robotics technologies for the enhancement of organisational processes and for their deployment and use in products and service, codes of professional ethics for IT professionals and managers in user organisations, and end-user guidelines.
3. Corporate responsibility policies and cultures that support ethical development and use of AI & robotics (aimed at both developers, deployers/users and support organizations)
4. National and international guidelines, standards and certification for ethical AI & robotics (aimed at governance and standards organisations; indirectly affecting developers, deployers/users and support organizations)
5. Policy and regulation to support ethical practices in AI & robotics (aimed at governance and standards organisations; indirectly affecting developers and deployers/users)



6. Education, training and awareness raising for the ethical and social aspects of AI & robotics (aimed at educators and the media)

In a concluding section of the report, the results of the study are summarized and future work towards further implementation is discussed.

Submitted to H2020, awaiting approval



List of tables

- **Table 1:** List of acronyms/abbreviations
- **Table 2:** Glossary of terms
- **Table 3:** SIENNA High-Level requirements

List of acronyms/abbreviations

Abbreviation	Explanation
AI	Artificial intelligence
EC	European Commission
R&D	Research and development
R&I	Research and innovation

Table 1: List of acronyms/abbreviations

Glossary of terms

Term	Explanation
Artificial Intelligence	The science and engineering of machines with capabilities that are considered intelligent (i.e., intelligent by the standard of <i>human</i> intelligence).
Big Data	Extremely voluminous data sets that require specialist computational methods to uncover patterns, associations and trends in them.
Data mining	The process of discovering patterns in large data sets involving database systems, statistical analysis, and intelligent methods such as machine learning.
Deep learning	An approach to machine learning that applies artificial neural networks with hidden layers and the backpropagation method, in combination with powerful computer systems and voluminous training data.
Ethics by Design	The systematic inclusion of ethical guidelines, recommendations and considerations into design and development processes.
Intelligent agent	An artificially created, autonomous entity that can perceive its environment by means of sensors, act upon this environment through the use of actuators, and direct its activities towards reaching goals.
Machine learning	A set of approaches within AI where statistical techniques and data are used to “teach” computer systems how to perform particular tasks, without these systems being explicitly programmed to do so.
Risk assessment	a systematic process of evaluating the potential risks that may be involved in a projected activity or undertaking.
Robotics	The field of science and engineering that deals with the design, construction, operation, and application of robots.
Robot	Electro-mechanical machines with sensors and actuators that can move, either entirely or a part of their construction, within their environment and perform intended tasks autonomously or semi-autonomously.

Table 2: Glossary of terms



1. Introduction

1.1 Background

This report has been developed within the SIENNA project, a European Horizon 2020-funded project on the ethical and human rights dimensions of emerging technologies.¹ A major focus of the SIENNA project is on the ethical and human rights aspects of AI and robotics. We have already performed extensive studies of ethical aspects of AI and robotics, the legal and human rights context for AI and robotics, existing ethical codes and guidelines for AI and robotics, the state of the art in AI and robotics and its social and economic impacts, and public awareness and acceptance of AI and robotics.² This is the first study in which we develop our own proposals. Based in part on our previous studies, we hereby propose an extensive ethical framework for the development and use of AI and robotics technologies.

1.2 Objectives

This report proposes a comprehensive strategy for ethical AI and robotics. That is, it proposes, at least in outline, a comprehensive set of methods and procedures for developing, deploying and using AI and robotics systems in a way that adheres to ethical principles. The strategy that we propose addresses all actors in society, particularly developers, deployers, users, regulators and educators. All have a role in bringing about ethical AI and robotics. Within this general strategy, we pay particular attention to methods and procedures for ethical research and innovation (R&I) in AI and robotics. Ethical R&I is often key for ensuring ethical standards for new technologies. In R&I, major decisions are made about what technological solutions to develop and which ones not to develop, and R&I often comes with prescriptions about deployment and usage as well. However, we will also pay attention to methods for ethical deployment and use, and to the role of organisations that market and use AI and robotics, technologies, as well as policy makers, regulators and educators, in bringing it about.

1.3 Structure of the report

The main body of the report consists of two parts after this introduction (section 1). Section 2, “A Strategy for AI and Robotics,” proposes an overall strategy for promoting ethical AI and robotics. It starts with an identification of relevant actors and six categories methods for obtaining ethical AI & robotics. It then proceeds to discuss the six categories of methods in more detail, and concludes with a section on how the methods can be developed (further) and how actors can be motivated to use them. Section 3, “A framework for Ethics by Design”, contains a detailed proposal for methods for incorporating ethical criteria into the design and development methodologies for AI and robotics. It first proposes a generic method for doing this, after which it contains a detailed discussion of doing it in relation to three popular development methodologies: CRISP-DM, Agile and the V-Model. In a

¹ See <https://www.sienna-project.eu/>.

² See reports D4.1, D4.2, D4.3, D4.5 and D4.6 at <https://www.sienna-project.eu/publications/>



concluding section (4), the results of the study are summarized and future work towards further implementation is discussed. Finally, in two annexes, detailed ethical guidelines are proposed for the incorporation of ethical criteria into Agile and the V-Model.

The role of ethical principles

It is not an objective of this report to develop or propose general ethical principles or guidelines for AI and robotics. By now, there is already enough convergence, in our opinion, on ethical principles for AI and robotics. Over the course of 2019, in particular, many countries and international organizations proposed general ethical guidelines for AI. Notably, 2019 saw the Ethics Guidelines for Trustworthy AI of the High-Level Expert Group on Artificial Intelligence (HLEG-AI, 2019), the Recommendation of the Council on Artificial Intelligence of the OECD (2019), the guidelines for Ethically Aligned Design from the Institute of Electrical and Electronics Engineers (IEEE, 2019), and the Governance Principles for a New Generation of Artificial Intelligence: Develop Responsible Artificial Intelligence China’s Ministry of Science and Technology (2019).

As several analysts have observed, there is a remarkable convergence between these recent sets of ethical guidelines. This was concluded, amongst others, in a recent study of the EU Horizon 2020—funded SHERPA project [FN], which was co-authored by some of the authors of this study (Ryan et al., 2019). The three main sets of guidelines (HLEG-AI, OECD and IEEE) display remarkable agreement in content, even though they have different formats and wordings. These documents are in essential agreement, it was found, on nine key ethical principles that include *privacy, autonomy, freedom, dignity, safety and security, justice/fairness, responsibility/accountability, well-being (individual, societal and environmental) and transparency*. In addition, none of these documents proposed major principles outside of this list. Even the Chinese guidelines converges remarkably with more “Western” guidelines: they by and large reflect these ethical principles as well.

1.4 Scope and limitations

In this report, as well as in future SIENNA proposals, we will adopt these nine key ethical principles as a starting point for ethical guidance. Specifically, given that this is a European Union funded project, we will adopt, with minor adaptations, the European version of these principles. That is, we will adopt the ethics guidelines for trustworthy AI of the HLEG-AI as our guiding set of principles, specifically its seven ethics requirements for trustworthy AI in which these nine principles are contained: *Human agency and oversight; Technical robustness and safety; Privacy and data governance; Transparency; Diversity, non-discrimination and fairness; Societal and environmental well-being; and Accountability*. Because of the strong similarities between these guidelines and others used outside the European union, we expect this study to have applicability outside the European Union as well.

These kinds of general guidelines will not be sufficient to offer ethical guidance for particular products and applications, or specific contexts of use. More detailed guidelines will also be needed to address such issues, for example, ethical guidelines for unmanned aerial vehicles, or for healthcare



applications of AI, or for predictive data analytics techniques. When needed, we will propose such more detailed guidelines. Our greatest concern in this report, however, is to operationalize ethical guidelines: how to make them directly usable by particular actors for particular practices. This is what much of this report will center on.

Particular attention will be paid to methods for the ethical development of AI & Robotics technologies. A large part of the report will be focused on such methods, under the heading of *Ethics by Design*. Section 3 of the main body of the report will be devoted to it, as well as the two annexes, which will develop Ethics by Design for two of the most often used development methodologies in AI and robotics.

Submitted to H2020, awaiting approval



2. A Strategy for Ethical AI and Robotics

As we argued, a set of ethical guidelines or principles is only one component of a strategy for ethical AI & robotics. It could provide some direction to activities, but only in a very general sense. Many more elements need to be in place to achieve the objective of ethical AI & robotics. Consider, for example, the development of AI & robotics technologies. Developers and other stakeholders involved, like most people, have certain ethical views and moral leanings that they respect. However, this may colour the development process. When given a list of ethical principles for AI, some developers may endorse them and make attempts to adhere to them in their activities. Such a set may point developers to actively focus on ethics during the development process. A set of principles, nevertheless, may not always be successful. Programmers could easily fail to do so due to either a lack of training in ethics, lack of knowledge of how to apply ethical principles in technology development, lack of support from management, lack of inclusion of ethics criteria in quality assessment frameworks or corporate social responsibility strategies, or other reasons. Much more is needed to make actors both motivated and competent in the incorporation of ethical considerations in their practices, and to support actors in collaborative practices towards this shared objective.

A sound strategy for ethical AI & robotics should in our view do three things:

- *Identify relevant actors*
- *Identify methods that these actors can use to contribute to ethical AI & robotics*
- *Propose ways in which these methods can be made available to these actors, and ways to motivate them to use them*

An overall strategy will be proposed in this report. Such a strategy is, in our view, a first step towards realizing ethical AI & robotics. A second step is the successful implementation of the strategy by relevant actors. Implementation will be a large part of the future focus of the SIENNA project, and future deliverables (particularly D5.4 and D6.6) will reflect this focus.

We will now proceed to identify the most relevant actor categories, and then propose relevant methods for each of them, including some shared methods that apply to different actor categories. We will end with a brief discussion of ways to make the methods available to actors and ways to motivate them.

Actors

The following actor categories are most relevant for our purposes. They have been selected on the basis of having the most influence on how AI & robotics technologies are developed, used, and perceived, and thereby on what their impacts and ethical aspects are:

- | |
|--------------------------------------------------------------------------|
| <i>1. AI & robotics developers</i> |
| <i>2. AI & robotics development support organizations</i> |
| <i>3. Organizations that deploy and use AI & robotics technology</i> |



<i>4. Governance and standards organizations</i>
<i>5. Educational and media organizations</i>
<i>6. Civil society organizations and the general public</i>

We will now discuss them in turn.

1. AI & robotics developers

Within this broad category, we can make some further distinctions. At the organizational level, developers include firms that develop AI & robotics technologies and research institutes (universities and other research performing organizations) that engage in research and innovation in AI & robotics. At the intra-organisational level, there are various units within these institutions that are involved in the planning, support and carrying out of R&I activities. At the individual level, there are also professionals in various roles (e.g., IT project manager, IT director, hardware technician, professor in robotics) that are actors in AI & robotics development.

2. AI & robotics development support organizations

These are organizations that provide support to the R&I activities of AI & robotics firms and research institutes. These include business and industry organisations (also known as trade organisations): organisations that support companies in a certain sector; chambers of commerce; research funding organisations; investment banks and other investors and funders; associations of universities and research institutes; science academies and associations of science academies; professional organisations for the AI & robotics fields; advisory and consultancy firms for companies and research institutes.

3. Organizations that deploy and use AI & robotics technology

These are private and public organisations that use AI & robotics. Its usage can be intended to improve or support various organizational functions, including operations, finance, marketing, human resources, customer service, and other. Within these organisations, one can furthermore define various units and professional roles associated with the deployment and use of AI systems within or by the organization, such as information technology managers, database administrators, and development operations engineers. Note that some organizations are simultaneously developers and users of AI & robotics systems. For example, tech companies like Apple and Google develop AI technologies, but also use them within their own organization.

4. Governance and standards organisations

These are organisations involved in developing, implementing or enforcing policies, standards and guidelines, specifically those regarding the development, deployment and use of AI & robotics technologies. It should be noted that organizations also make policies and guidelines for themselves. These are not our concern here. This category is intended to refer to organizations that develop or implement guidelines, policies, regulations and standards for others. This includes, first of all, national, local and supranational governments, as well as government-instituted or -supported advisory and



regulatory bodies. They also include intergovernmental organisations like the United Nations, the Council of Europe, and the World Health Organization. Also included in this category are national and international standards, certification, quality assurance, accreditation and auditing organisations. Policies, standards and guidelines can also be issued by many of the AI & robotics development support organisations discussed earlier.

5. Educational and media organisations

Educational institutes and media organisations both have a significant role, albeit a quite different one, in shaping people's knowledge and understanding of AI & robotics, the ethical issues associated with them, and the ways in which these ethical issues can be addressed. Educational organisations, from elementary school to postgraduate education, provide the major vehicle by which individuals acquire knowledge, skills and insights regarding AI & robotics, their impacts on society, their ethical aspects, and ways to address ethical issues in their profession. Of course, not only educational organisations provide education and training. Companies may, for example, organize their own in-house trainings as well. Media organisations have a large role in generating public awareness and understanding of AI & robotics and the ethical issues raised by them and therefore should also be recognized as actors with respect to ethical AI & robotics.

6. Civil society organisations and the general public

Civil Society Organisations (CSOs) are non-governmental, not-for-profit organisations that represent the interests and will of citizens. They may be based on cultural, political, ethical, scientific, economic, religious or philanthropic considerations. They include civic groups, cultural, groups, consumer organisations, environmental organisations, religious organisations, political parties, trade unions, professional organisations, non-governmental policy institutes, activist groups, and several other kinds. Many CSOs want to have a role in public policy or influence the way that organizations function in which they have an interest. For some of them, the development and use of AI will be a concern, and as a result, these organisations will function as agents with respect to public policy and the actions of relevant other organisations. The general public, finally, can also perform as an actor, through its public opinions, voting patterns, consumer purchases, and use or nonuse of AI & robotics products and services.

Finally, it is worth mentioning that amongst and within these various kinds of organisations and units, there are also those that have a specific focus on ethics. These include ethics research units, ethics policy units, ethics officers, research ethics committees, integrity offices and officers, corporate social responsibility units and officers, ethics educational programmes, ethics advisory bodies, and national and international ethics committees. However, ensuring ethical standards and practices is not only the responsibility of such organisations and units; all of the listed actors have such responsibilities, although ethics organisations and units will often have a special role in ensuring the proper inclusion of ethics concerns in practices.



Methods

In the context of this report, methods are means by which actors can implement ethical guidelines and considerations. Our identification of methods for ethical AI & robotics builds on earlier proposals of the HLEG-AI (2019) and IEEE (2019). Both reports propose methods for the implementation of ethical guidelines in relation to different actors. The HLEG makes a distinction between what they call technical and non-technical methods, both of which apply to all stages of the development and use lifecycle of AI systems. Technical methods include ethics by design methods, explanation methods for transparency, methods of building system architectures for trustworthiness, extensive testing and validation, and the definition of quality of service indicators. Non-technical methods include regulation, codes of conduct, standardization, certification, accountability via governance frameworks, education and awareness to foster an ethical mindset, stakeholder participation and social dialogue, and diverse and inclusive design teams.

The IEEE (2019) report has a chapter on “methods to guide ethical research and design” for researchers, technologist, product developers and companies (pages 124-139), and a chapter on policies and regulations by governing institutions and professional organizations (pages 198-210). In its methods for ethical R&D chapter, it considers both individual and structural approaches, and distinguishes between three overall approaches: interdisciplinary education and research, corporate practices on AI & robotics, and responsibility and assessment. In its policy chapter, the IEEE advocates methods such as the founding of national policies and business regulations for SIS on human rights approaches, the introduction of support structures for the building of governmental expertise in AI and robotics, and the fostering of AI & robotics ethics training in educational programs.

The methods proposed by the HLEG-AI and IEEE are partially overlapping and in part complementary. Drawing from them, we propose six sets of methods for the ethical development and use of AI & robotics³, for the different classes of actors that were defined earlier:

1. Methods for incorporating ethics into research and development of AI & robotics (aimed at AI & robotics developers and support organizations)
2. Methods for incorporating ethics into the deployment and use of AI & robotics (aimed at organisations that deploy and use AI & robotics technology)
3. Corporate responsibility policies and cultures that support ethical development and use of AI & robotics (aimed at both developers, deployers/users and support organizations)
4. National and international guidelines, standards and certification for ethical AI & robotics (aimed at governance and standards organisations; indirectly affecting developers, deployers/users and support organizations)
5. Education, training and awareness raising for the ethical and social aspects of AI & robotics (aimed at educators and the media)

³ Points 1, 3-6 are taken from the SHERPA development and use guidelines (Brey, Lundgren, Macnish and Ryan, 2019). Point 2 is an added point.



6. Policy and regulation to support ethical practices in AI & robotics (aimed at governance and standards organisations; indirectly affecting developers and deployers/users)

We will refrain, for now, to propose methods for CSOs and the general public, taking into account that their role in ethical AI & robotics is often more indirect. We will now discuss these sets of methods in some more detail and relate them to the roles and responsibilities of different actors.

Methods for incorporating ethics into research and development

These are methods for making ethical considerations, principles, guidelines, analyses or reflections part of research and development processes. They apply to the first actor category identified above: AI & robotics developers. Four main classes of methods fall into this category:

- | |
|---------------------------------------------------------------------------------------------------------|
| <i>1. Research ethics guidelines and protocols for R&I in AI & robotics</i> |
| <i>2. Ethical impact assessment methodologies for emerging AI & robotics</i> |
| <i>3. Ethics by design methodologies for AI & robotics</i> |
| <i>4. Codes of professional ethics for researchers and developers of AI & robotics technologies</i> |

We will now discuss them in turn.

1. Research ethics guidelines and protocols for R&I in AI & robotics

Research ethics guidelines and protocols for AI & robotics are ethics guidelines and procedures by which researchers, developers, research ethics committees and ethics officers can ethically assess R&I proposals and ongoing R&I practices. Such ethical assessments may or may not be accompanied with specific recommendations to proceed differently. They can, in either case, be used to improve R&I plans and practices so as to make them more ethical. As of the moment of publication of this report, few research ethics guidelines and protocols specifically for AI and robotics were in existence (see our report D4.3 Survey of REC approaches and codes for Artificial Intelligence & Robotics). While there is an abundance of general ethical guidelines for AI and robotics, few specifically focus on R&I practices and on the role of research ethics committees. We are currently working on our own proposal for research ethics guidelines and protocols for AI & robotics, and will present them in a future SIENNA report.

2. Ethical impact assessment methodologies for emerging AI & robotics

Ethical impact assessment methodologies are methods for assessing present and potential future impacts of emerging technologies, including specific products and applications, and identifying ethical issues associated with these impacts. EIA, in short, is an approach for assessing not only present but also potential future ethical issues in relation to a technology. EIA, in its current form, was developed within the EU FP7 SATORI project [FN]. It has also been developed into a CEN standard (CEN, 2017). EIA is not just a method for AI & robotics developers, but can also be used, amongst others, by



governments in order to support technology policy, and by research funding organisations to help set priorities in research funding.

3. Ethics by design methodologies for AI & robotics

Ethics by design methodologies for AI & robotics are methods for incorporating ethical guidelines, recommendations and considerations into design and development processes. They fill a gap that exists in current research ethics approaches, which is that it is often not clear for developers how to implement ethical guidelines and recommendations, which are often of a quite general and abstract nature. Ethics by design methodologies identify how at different stages in the development process, ethical considerations can be included in development, by finding ways to translate and operationalize ethical guidelines into concrete design practices. Ethics by design approaches have been in existence in computer science and engineering since the early 1990s, initially under the name Value-sensitive design (Friedmann Kahn & Borning, 2006) and later also under the label of Design for Values (Van den Hoven, Vermaas and Van de Poel, 2015). In recent years, the term “ethics by design” has come into vogue. Recently, an extensive ethics by design approach for AI was published as part of the EU Horizon 2020-funded project SHERPA (Brey, Lundgren, Macnish and Ryan, 2019). As far as we can see, no other full-blown ethics by design approaches have yet been published for AI & robotics, although the IEEE is working on one. In this report, we build on the SHERPA report to present an extended approach for ethics by design that has wider applicability than the one proposed in that report.

4. Codes of professional ethics for researchers and developers of AI & robotics technologies

Codes of professional ethics, also called codes of conduct, are codified personal and corporate standards of behaviour that are expected in a certain profession or field. These codes are often set by professional organisations. To our knowledge, no internationally accepted codes of ethics for either artificial intelligence specialists or robotics engineers are currently in existence, and few if any national codes for these professions exist either. Wider codes of ethics, for computer scientists and electrical engineers, are in existence and cover the AI and robotics professions as well. However, these broader codes do not address the specific challenges and responsibilities of AI and robotics specialists. In this report, we do not attempt to propose codes of professional ethics for these professions. We could make some initial proposals, however, in later studies in the SIENNA project.

In the HLEG and IEEE reports, various other methods for incorporating ethics into R&D are mentioned. Some of these can however, in our opinion, be subsumed under ethics by design approaches. These include, amongst others, the development and use of explanation methods for transparency, extensive testing and validation, the definition of quality of service indicators, and better technical documentation. Others will be discussed under the heading of “corporate social responsibility cultures” below. One method merits special attention, however: interdisciplinary research, which is proposed in the IEEE report. Interdisciplinary research is, in our view, an important component of ethical AI & Robotics, if it involves collaborations that bring engineers and scientists into contact with social science and humanity scholars, including ethicists. Such research activities allow for a better incorporation of social and ethical concerns into engineering practice, and are therefore highly advisable, at different stages of the R&D continuum.



Methods for incorporating ethics into the deployment and use of AI & robotics

After the development of AI & robotics systems, services and solutions, they are deployed by organisations or individuals in order to be used.⁴ The deployment and use of these technologies often require their own ethical guidelines and solutions, that are to some extent different from those that apply to their development. Ethical questions that are typically asked in relation to deployment and use include questions like: Is it ethical to deploy a system that is intended to do X / is capable of doing X / can be used to do X? How can unethical uses of the system be monitored and prevented? What is the responsibility of different actors in preventing or mitigating unethical use? What policies to prevent unethical use should be put in place and how can they be implemented effectively?

Deployment and use scenarios come in various forms, but the following are the most typical:

- (1) Deploying AI or robotics technology to enhance organisational processes. An organisation acquires AI or robotics technology, and uses it within its own organisation to improve organisational processes such as manufacturing, logistics, and marketing. End-users are IT specialists or other employees in the organisation.
- (2) Embedding AI and robotics technology in products and services. An organisation acquires AI or robotics technology, and incorporates it into products or services that it offers to customers. This is a different application of AI and robotics than its application in the development, manufacturing and marketing of products and services. For example, AI can be used to better design, manufacture or market automobiles that themselves do not contain AI technology. AI and robotics technologies can be embedded in products and services for different purposes:
 - a. To enhance the value of a product or service for customers by offering enhanced functionality or usability. E.g., by powering an online dating service with AI algorithms, or by enhancing an automobile with a self-drive mode.
 - b. To enhance the value of a product or service through intelligent monitoring, self-repair, communications with customer service, or data collection for future upgrades.
 - c. To further the interests of the organisation or of third parties, for example, by collecting data for marketing purposes or allowing for targeted messaging.

It is not always clear who is the end-user of the AI and robotics technology in these three scenarios, since the end-user of AI or robotics technology embedded in a product or service may be different from the end-user of that product or service, and there may also be multiple end-users (e.g., Uber drivers and customers using the same AI algorithms).

Taking these scenarios into consideration, the following four methods can contribute to ethical deployment and use of AI & robotics technologies:

⁴ Of course, deployment and use cycles are often followed by repeated redevelopment of systems.



- (1) Operational ethics guidelines and protocols for the deployment and use of AI and robotics technologies for the enhancement of organisational processes
- (2) Operational ethics guidelines and protocols for the deployment and use of AI and robotics technologies in products and services
- (3) Codes of professional ethics for IT managers, technical support specialists and other management, IT and engineering staff responsible for the deployment and use of the AI & robotics technologies in an organisation or its embedding in products and services
- (4) End-user guidelines for ethical usage of (products and services that include) AI and robotics technologies

In Brey, Lundgren, Macnish and Ryan (2019), the previously mentioned SHERPA report, proposals were made for the first and, to some extent, the second of these methods. Building on two widely used models for the management and governance of information technology in organisations, ITIL and COBIT, as well as on the ethics requirement of the High-Level Expert Group on AI, this report proposed operational guidelines for the deployment and use of AI systems (including AI-powered robotic systems) in organisations. We will not do further work on these guidelines in this report. We also will not attempt to further develop codes of professional ethics for the different professions responsible for the deployment and use of AI & Robotics technologies. Often, codes of ethics will be in place for these professions, but they might need updates to take into account the specific demands imposed by AI & robotics technologies. We also will not attempt to develop (generic) guidelines for end-users in the context of this report.

Corporate responsibility policies and cultures

Ethical guidelines and professional ethical codes, even when fully operationalized for particular practices, will have little impact if they are not supported by organisational structures, policies and cultures of responsibility. In Brey, Lundgren, Macnish and Ryan (2019), specifically the division of the report with guidelines for the ethical deployment and use of AI (p. 53-87), an attempt was made to include these wider considerations of responsibility in organisations in the guidelines that were proposed. For instance, requirement 1 in this report, which targets the board of directors of companies, reads as follows:

Requirement 1. The board of directors should direct in its IT governance framework that IT management adopts and implements relevant ethical guidelines for the IT field, and should monitor conformity with this directive. There should be an appointed representative at each level of the organisation, including the board of directors, who are 'ethics leaders' or 'ethics champions', and who should meet regularly to discuss ethical issues and best practice within the organisation. The ethics leader from the board of directors should be responsible for the ethical practice of the whole organisation (p. 61).

Requirements 2, 3 and 4, which targets IT management, are as follows:



Requirement 2. The IT management strategy should include the adoption and communication to relevant audiences of ethics guidelines for AI and big data systems, define corresponding ethics requirements within role and responsibility descriptions of relevant staff, and include policies for the implementation of the ethics guidelines and monitoring activities for compliance and performance (p. 64).

Requirement 3: The IT management strategy should include the design and implementation of training programs for ethical awareness, ethical conduct, and competent execution of ethical policies and procedures, and these programs should cover the ethical deployment and use of the system. More generally, IT management should encourage a common culture of responsibility, integrating both bottom-up and top-down approaches to ethical adherence (p. 64-65).

Requirement 4: Consider how the implementation of the AI and big data systems ethics guidelines, and other IT-related ethics guidelines, affects the various dimensions of IT management strategy, including overall objectives, quality management, portfolio management, risk management, data management, enterprise architecture management, stakeholder relationship management. Ensure proper adjustment of these processes. There will be different levels of risk involved, depending upon the application, so the levels of risk need to be clearly articulated to allow different responses from the organisation's ethical protocols (p. 65).

These guidelines, and several others that are proposed, serve as meta-guidelines for the proper implementation of ethics guidelines for AI & robotics in organizations. They point out that proper implementation of ethics considerations in organizations involves much more than the development and distribution of operationalized ethics guidelines, but also requires leadership from the top, adjustment of existing management strategy, definitions of roles and responsibilities, training of staff, monitoring and assurance activities, and encouragement of a common culture of responsibility. While these guidelines were developed for organisations that deploy and use AI & robotics technologies, they are also applicable to organizations that engage in AI & Robotics R&D.

National and international guidelines, standards and certification

In this report, we distinguish between *operational ethics guidelines*, which are detailed, practical guidelines developed for specific practices by specific actors, and *general ethics guidelines*, which are statements of ethical principles and general guidelines that apply to a broad range of actors and practices. While it is possible to develop operational guidelines without general guidelines, it is often beneficial to have shared general guidelines on the basis of which operational guidelines are developed. These guidelines can be supported by national governments and intergovernmental organisations. Currently the two most important sets of international guidelines for AI & robotics technologies are the Recommendation of the Council on Artificial Intelligence of the OECD (2019) and the Ethics Guidelines for Trustworthy AI of the High-Level Expert Group on Artificial Intelligence of the European Commission (HLEG-AI, 2019). These two documents currently serve as the two most important international guidance documents for ethical issues in AI & robotics.



Next to such general guidelines, which are directed at all actors, there are also ethical guidelines that are general rather than operational, but that are focused on specific actors or practices. The guidelines for Ethically Aligned Design from the Institute of Electrical and Electronics Engineers (IEEE, 2019) are a case in point. These specifically apply to design practices, and are of greatest relevance to technology developers.

Standards, developed by recognized national and international standards organisations or by particular (associations of) companies or organisations, are different from ethics guidelines in two ways. First, they apply to specific products, services, processes or methods, while ethics guidelines apply to any action, thing or event that has ethical implications. Second, they define specific norms or requirements to which the phenomenon to which the standard applies must adhere. Standards are intended to leave limited room for subjectivity and interpretation, and are intended to define intersubjective requirements that different actors can apply, identify or assess.

Standards sometimes aim to codify ethical requirements, procedures or methods. Examples are ISO 26000, which is an international standard for corporate social responsibility, CEN CWA 17145-1, which is a standard for ethics assessment by research ethics committees, and CEN CWA 17145-2, which is a standard for the method of ethical impact assessment for R&I. Standards can also *include* ethical requirements, procedures or methods, while not themselves having ethics as a focus. For example, ethics is discussed in the context of the ISO 9000 and 9001 standards for quality management.

For AI & robotics, a remarkable number of ethical standards are currently being developed by IEEE as part of its Ethically Aligned Design programme (IEEE, 2019). A total of 13 standards are in development, including standards for ethics by design, transparency of AI systems, algorithmic bias, data privacy, ethically driven robotics and automation systems, and automated facial analysis technology. ISO also has several standards in development that focus in part or in whole on ethical issues, including standards for identifying ethical and societal concerns in AI systems, bias in AI systems, trustworthiness of AI systems, quality assurance in AI and risk assessment in AI.

Certification is the process by which an external third party (typically a certifying body) verifies that an object, person or organization is in possession of certain characteristics or qualities. Amongst others, certification can be applied to persons, in professional certification, to products, to determine if it meets minimum standards, and to organizations or organizational processes, through external audits, to verify that they meet certain standards. Certification can be a means to verify and validate the quality of ethics processes and procedures in organisations. In relation to standards, in particular, certification can be a means of ensuring conformity to the requirements of the standard. IEEE is currently developing its own certification programme to certify adherence to the ethics standards it is developing. ISO does not do certification itself, but third-party certification organisations could in the future assess compliance to ISO ethics-related standards for AI.

Education, training and awareness raising

Education is a powerful method for stimulating ethical behaviour in relation to AI & robotics. In professional and academic education, specifically, education that concerns ethical and social issues in



AI & robotics would benefit future professionals, especially those in the AI & robotics field, but also those in other fields who may deploy and use these technologies in the future. Given the seriousness of ethical issues in the AI & robotics fields, a required ethics course for AI and robotics students seems advisable. Such a course could cover key ethical issues in AI & robotics, ethical guidelines and their application, responsibilities of AI and robotics professionals, and relevant standards, laws, policies, and approaches for ethical AI & robotics. Methodologies for ethics by design could be part of such a course, but for these to be used by future professionals in actual design practice, it might be better if these were to be incorporated in the standard design methodologies used in these fields.

Most professionals who develop and use AI & robotics did not have ethics education in these areas in their professional education. For them, continuing education programmes that include ethics of AI and/or robotics would be valuable. Such training programmes could even be accompanied by professional certification, for example, certification in ethics by design methodology, algorithmic bias avoidance, preparing for ethics review, or all-round ethical practice in AI or robotics. Next to external organisations setting up such training and education programmes, organisations could of course also organize their own in-house training in ethics for AI & robotics.

Turning now from educational institutions to the media, we should acknowledge that media organisations have a large role in generating public awareness and understanding of AI & robotics, including the ethical issues raised by them. These are complicated technologies that are difficult to understand for the average person. Since they are expected to have major impacts on people's lives, a proper understanding of them and the ethical issues they raise is important, and media companies are the most important type of organization who can provide such an understanding to the general public. Therefore, relevant media stories on AI & robotics and its social and ethical dimensions, whether in print, podcast, television or other formats, are important. While media organisations have a major responsibility here, AI & robotics developers also have a responsibility to communicate with the public about these issues, and governments in ensuring that sufficient information is provided.

Policy and regulation

While policy can be made by any kind of organization, our concern here is with public policy, as made by governments, as well as the laws and regulations issued by them. The key question here is: what policies, laws and regulations should governments develop, if any, to stimulate the ethical development, deployment and use of AI & robotics? Policies, laws and regulations can relate to ethical criteria in three ways: they can explicitly institute, promote or require ethics guidelines, procedures, or bodies; they can have a focus on upholding certain moral values or principles without explicitly identifying them as ethical (e.g., well-being, privacy, fairness, sustainability, civil rights); and they either explicitly or implicitly take on board ethical considerations in broader social and economic policies.

Governments are currently at a decision point for AI & robotics policy. What should they do, and how can they avoid regulating too little as well as regulating too much? Decisions that relate to ethics include the following:



- Whether or not to issue, or support the issuing of, ethical guidelines for AI & robotics
- Whether or not to put any ethical guidelines for AI & robotics into law
- Whether or not to revise existing institutional structures to better account for ethical issues or to create new governmental bodies or unites for ethical and social issues in AI & robotics
- Whether or not to mandate ethics standards, certification, education, training, ethical impact assessments or ethics by design methods in relation to ethics of AI & robotics
- Whether and how to introduce new legislation and regulations to for morally controversial AI & robotics technologies, such as automated tracking, profiling and identification technologies, behaviour and affect recognition technologies, and automated lethal weapons
- How to include ethical considerations concerning AI & robotics in policies, laws and regulations, both ones that pertain to AI & robotics specifically and more general ones that need to be updated to account for AI & robotics, such as in the areas of consumer protection, data protection, criminal law, non-discrimination provisions, civil liability and accountability
- What financial support and funding to provide, if any, for ethics research, ethics education, ethics dialogue, ethics awareness raising and other ethics initiatives in relation to AI & robotics
- How to regulate the government's own use of AI & robotics so as to ensure ethical conduct

See also the forthcoming SIENNA report D5.6, *Recommendations for the enhancement of the existing EU and international legal framework*, which will contain our proposals for new EU and international legislation and regulations to support ethical AI & robotics.

Finally, a general remark regarding these methods: it remains to be seen whether ethical AI & robotics are best served by specific ethics standards, certification, design methodologies, audits, policies and other methods, or whether it is better to integrate ethics concerns into broader standards, policies, audits, etc. This probably varies from situation to situation, but should receive proper attention as an issue to account for.

Making methods available and motivating actors

In the preceding discussion of methods, we have already made a number of suggestions regarding the responsibility of different actors for developing and making available different types of methods. Obviously, governments are the responsible party for the development governmental policies, laws and regulations, and universities are the ones that would development of ethics courses in degree programmes in AI and robotics. In other cases, it may not be immediately obvious which actor would be responsible for developing and advocating for a particular method. Which actor would be responsible for developing methods of ethical impact assessment, for example, or for developing operational ethics guidelines for the deployment and use of AI in organisations? Often, this is a matter of particular actors stepping up and taking on such responsibilities. It was not written in stone that the IEEE should embark on in an extensive programme to develop ethical guidelines, methods, standards



and certification for the design and deployment of AI and robotics systems, but it nevertheless chose to do so.

On the other hand, actors may fail to step up, leaving a responsibility vacuum in society due to which important methods for ethical AI & robotics are not being developed and implemented. If this is to occur, then governments are often seen as the responsible actor to step in and enact policies, laws and regulations that help fill this vacuum. Governments, after all, have a particular responsibility for promoting the public good, protecting individual rights, and supporting fair socioeconomic conditions, and also have powers to stimulate and compel other actors to act responsibly and in the public interest.

3. Conclusion

The aim of this report was to propose a comprehensive strategy for ethical AI and robotics. In addition, it was an aim to present an approach for Ethics by Design, as part of that strategy. These two aims were undertaken in two major sections of the report, “A strategy for Ethical AI and Robotics” (section 2) and “A framework for Ethics by Design” (section 2).

In section 2, it was claimed that a strategy for ethical AI and robotics should contain three components: (1) an identification of relevant actors; (2) an identification of methods that these actors can use to contribute to ethical AI & robotics, and (3) proposals of ways in which these methods can be made available to these actors, and ways to motivate them to use them. Subsequently, these three components were given content in the report. Six main classes of relevant actors were defined, including AI & robotics developers; AI & robotics development support organizations; organizations that deploy and use AI & robotics technology; governance and standards organizations; educational and media organizations; and civil society organizations and the general public.

Next, six types of methods for ethical AI & robotics were discussed and related to these classes of actors: methods for ethical development and design, methods for ethical deployment and use, corporate responsibility policies and cultures, national and international guidelines, standards and certification, policy and regulation actions (by governments), and education, training and awareness raising. Finally, it was briefly discussed how these methods can be made available to actors.

Our discussion of methods for ethical AI & robotics in section 2 is only brief, and we did not have the room to arrive at detailed proposals for many of the methods that we discuss. For many of the proposed methods, however, we refer to both past and planned deliverables that we have completed or are preparing within the SIENNA project, or to other initiatives in which these methods have been or are being developed.

As stated earlier, this strategy is only a first step towards ethical AI & robotics, and a second step consists of its implementation. This requires both the further specification and operationalisation of the methods described in it, the mobilisation of stakeholders and the implementation of the strategy



together with these stakeholders. This is what we will spend much of the remainder of the SIENNA project on.

Submitted to H2020, awaiting approval



References

- Brey, P., Lundgren, B., Macnish, K. and Ryan, M. (2019). *Guidelines for the development and use of SIS*. Deliverable D3.2 of the SHERPA project. <https://doi.org/10.21253/DMU.11316833>.
- CEN (2017). *Ethics assessment for research and innovation - Part 2: Ethical impact assessment framework*. CEN workshop agreement, CWA 17145-2.
- Friedman, B., Kahn, P. and Borning, A. (2006). 'Value Sensitive Design and Information Systems,' in *Human-Computer Interaction in Management Information Systems: Foundations* (eds. P. Zhang and D. Galletta). Armonk, NY: M.E. Sharpe.
- HLEG-AI (High-Level Expert Group on Artificial Intelligence) (2019). *Ethics Guidelines for Trustworthy AI*. Downloaded on 8-3-2020 at <https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines#Top>
- Hoven, Jeroen van den, Pieter E. Vermaas, and Ibo van de Poel, eds. (2015). *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains*. Springer Netherlands.
- IEEE (The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems) (2019). *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*, First Edition. IEEE. <https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/>
- Jansen, P., Brey, P., Fox, A., Maas, J., Hillas, B., Wagner, N., Smith, P., Oluoch, I., Lamers, L, Van Gein, H., Resseguier, A., Rodrigues, R., Wright, D. and Douglas, D. (2019). *Ethical Analysis of AI and Robotics Technologies*. D4.4 of the SIENNA project. <https://www.sienna-project.eu/publications/>.
- Ministry of Science and Technology of China (2019). *Governance Principles for a New Generation of Artificial Intelligence: Develop Responsible Artificial Intelligence*. A translation can be found at: <https://perma.cc/V9FL-H6J7>.
- OECD (2019). *Recommendation of the Council on Artificial Intelligence*. Retrieved on 8-3-2020 at <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>.
- Ryan, M., Philip Brey, Kevin Macnish, Tally Hatzakis, Owen King, Jonne Maas, Ruben Haasjes, Ana Fernandez, Sebastiano Martorana, Isaac Oluoch, Selen Eren, and Roxanne Van Der Puil (2019). *Ethical Tensions and Social Impacts*. Deliverable D 1.4 of the SHERPA project. <https://doi.org/10.21253/DMU.8397134>